Development and Performance of a Scalable Version of a Nonhydrostatic Atmospheric Model

A. A. Mirin and G. Sugiyama
Lawrence Livermore National Laboratory

S. Chen, R. M. Hodur, T. R. Holt, and J. M. Schmidt
Naval Research Laboratory

DoD HPC Users Group Conference 2001
18-21 June 2001

Outline

Development and Performance of a Scalable Version of a Nonhydrostatic Model

- **What is COAMPS?**
  - Definition
  - Operations
- **Present and Future Computer Resources**
- **Development of Scalable COAMPS:**
  - Background
  - Organization of Workload
  - Program Structure
  - Pre-processing/Analysis
  - Forecast Model
    - Domain Decomposition
    - Nesting
  - Test Results
- **Future Plans and Conclusions**

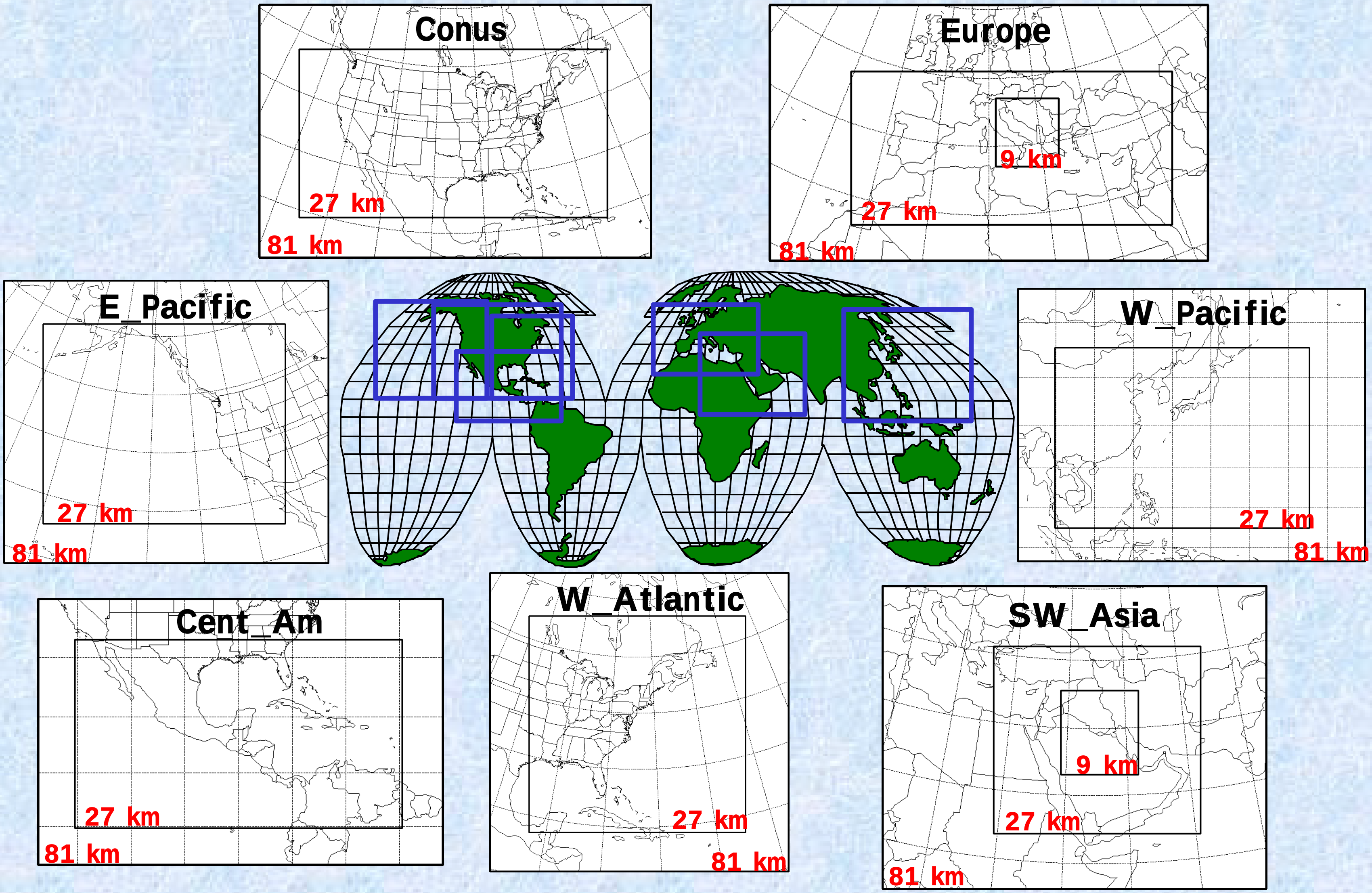
COAMPS

Coupled Ocean/Atmosphere Mesoscale Prediction System: **Atmospheric Components**

- **Complex Data Quality Control**
- **Analysis:**
  - Multivariate Optimum Interpolation Analysis (MVOI) of Winds and Heights
  - Univariate Analyses of Temperature and Moisture
  - 2D OI Analysis of Sea Surface Temperature
- **Initialization:**
  - Hydrostatic Constraint on Analysis Increments
  - Digital Filter
- **Atmospheric Model:**
  - Numerics: Nonhydrostatic, Scheme C, Nested Grids, Sigma-z, Flexible Lateral BCs
  - Parameterizations: PBL, Convection, Explicit Moist Physics, Radiation, Surface Layer
- **Features:**
  - Globally Relocatable (5 Map Projections)
  - User-Defined Grid Resolutions, Dimensions, and Number of Nested Grids
  - 6 or 12 Hour Incremental Data Assimilation Cycle
  - Can be Used for Idealized or Real-Time Applications
  - Single Configuration Managed System for All Applications
  - Operational at FNMOC:
    - 7 Areas, Twice Daily, using 81/27/9 km or 81/27 km grids
    - Forecasts to 72 hours
  - Operational at all Navy Regional Centers (w/GUI Interface)

COAMPS Operational Areas at FNMOC

As of September 8, 2000



COAMPS

Coupled Ocean/Atmosphere Mesoscale Prediction System: **Ocean Components**

- **Data Quality Control**
- **Analysis:**
  - 2D Multivariate Optimum Interpolation Analysis (MVOI) of Sea Surface Temperature on All Grids
  - 3D MVOI Analysis of Temperature, Salinity, Surface Height, Sea Ice, and Currents
- **Ocean Model:** Navy Coastal Ocean Model (NCOM)
  - Numerics: Hydrostatic, Scheme C, Nested Grids, Hybrid Sigma/z
  - Parameterizations: Mellor-Yamada 2.5
- **Features:**
  - Globally Relocatable (5 Map Projections)
  - User-Defined Grid Resolutions, Dimensions
  - Can be Used for Idealized or Real-Time Applications
  - Single Configuration Managed System for All Applications
  - Loosely coupled to COAMPS atmospheric model

Present and Future Computer Resources

- **Operations at FNMOC:**
  - Current: Cray c90 [16-processor (1), 8-processor (1)]
  - Sep 2001: SGI o3k [128 processor (1), 512 processor (1)]
- **Operations at Regional Centers:** SGI o2k [4-processor (1)]
- **Operations at DoE NARAC:** DEC [4-6 processors (1)];

  NARAC: National Atmospheric Release Advisory Capability
- **Research at NRL/DoD HPC Centers:**
  - SGI o2k [64-processor (1), 128-processor (3)]
  - SGI o3k [128-processor (1), 256-processor (5)]
  - DEC [8-processor (1)]
  - IBM [512-processor (1), 1200-processor (1)]
  - Cray T3E [544-processor (1), 1088-processor (1)]
  - Cray SV1 [16-processor (4), 24-processor (1)]
- **Research at LLNL:**
  - TeraCluster2000 [DEC 512-processor (1)]
  - IBM [512-processor (1)]

Scalable COAMPS

Background

- **COAMPS Original Design for Shared Memory Systems:**
  - 1980's: Cyber 205 [Vectorization]
  - 1990's: Multi-Processors (e.g., c90) [Multi-tasking]
- **New Scalable Architecture for FNMOC/HPC/LLNL:**
  - Hardware does not support vectorization
  - Necessitates new programming model:
    - Node to node communications (Message Passing Interface, MPI)
    - Processor to processor (MPI or OpenMP)
- **Complications:**
  - Domain Decomposition:
    - Overhead for developers
    - Complicates "non-local" processes
    - MPI is an evolving standard
  - FORTRAN Compilers:
    - Buggy
    - Different options/versions on different platforms
  - Few Development Tools (but getting better)

Scalable COAMPS

Organization of Workload: Joint NRL-LLNL Development

- **LLNL** (Art Mirin, Gayle Sugiyama):
  - Previous experience w/MOM, UCLA GCM
  - Focus on: Domain decomposition, Communications
  - Availability of DoE hardware: DEC, IBM
  - MOA w/NRL
- **NRL** (Jerry Schmidt, Teddy Holt, Sue Chen)
  - Focus on: Physics, I/O, Nesting, Pre-processing, Test suite
  - Availability of HPC hardware: T3E, O2K, IBM
  - Requirements for new FNMOC and HPC hardware
- **Development on:**
  - LLNL: DEC, IBM
  - NAVO: T3E, O2K
  - NRL DC, ARL: O3K
  - FNMOC: O2K

COAMPS Program Structure

Atmospheric Components

Pre-Processing/Analysis
(coama)

- Construct “data” record
- Generate grid information
- Generate surface fields
- Construct SST OI analysis
- Construct atmospheric MVOI analyses*
- Generate lateral boundary condition data for coamm from NOGAPS fields

Forecast Model
(coamm)

- Merge analysis increments and previous forecast fields
- Initialization
- Model integration*
- Output:
    - Pressure levels
    - Height levels
    - Surface fields
    - Sigma levels
    - Individual points

*Most time-consuming portion of job

COAMPS

Pre-processing/Analysis (coama)

•Shared Memory Structure:

- Arrays organized in i-, j-, k- structure
- Many i,j loops combined into one i-loop for vectorization
- Cray/SGI multi-tasking instructions for k-loops and MVOI volume loops
- Bicubic splines for staggering and de-staggering winds

•Distributed Memory Structure:

- Retain shared memory constructs
- Retain i-, j-, k- structure
- Use OpenMP for k-loops and MVOI volume loops
- Bicubic splines for staggering and de-staggering winds
- SST analysis moved to separate program

COAMPS

Forecast Model (coamm)

•Shared Memory Structure:

- Arrays organized in i-, j-, k- structure
- Many i,j loops combined into one i-loop for vectorization
- Cray/SGI multi-tasking instructions:
  - Outer (k-) loop for dynamics (levels)
  - j-loop for physics (vertical slabs)
- Bicubic splines for staggering and de-staggering winds

•Distributed Memory Structure:

- Retain i-, j-, k- structure
- Implement MPI for communications and MPI I/O for output
- Domain decomposition in x-, y- directions (user-defined for each nest)
- Arbitrary number of halo rows/columns (user-defined)
- Allow for OpenMP multi-tasking instructions:
  - Outer (k-) loop for dynamics (levels)
  - j-loop for physics (vertical slabs)
- Retain option for vectorization (i.e., collapsed loops)
- Use bilinear interpolations for staggering/de-staggering winds
- Drop unused code (e.g., simplified physics)

Vectorization in COAMPS

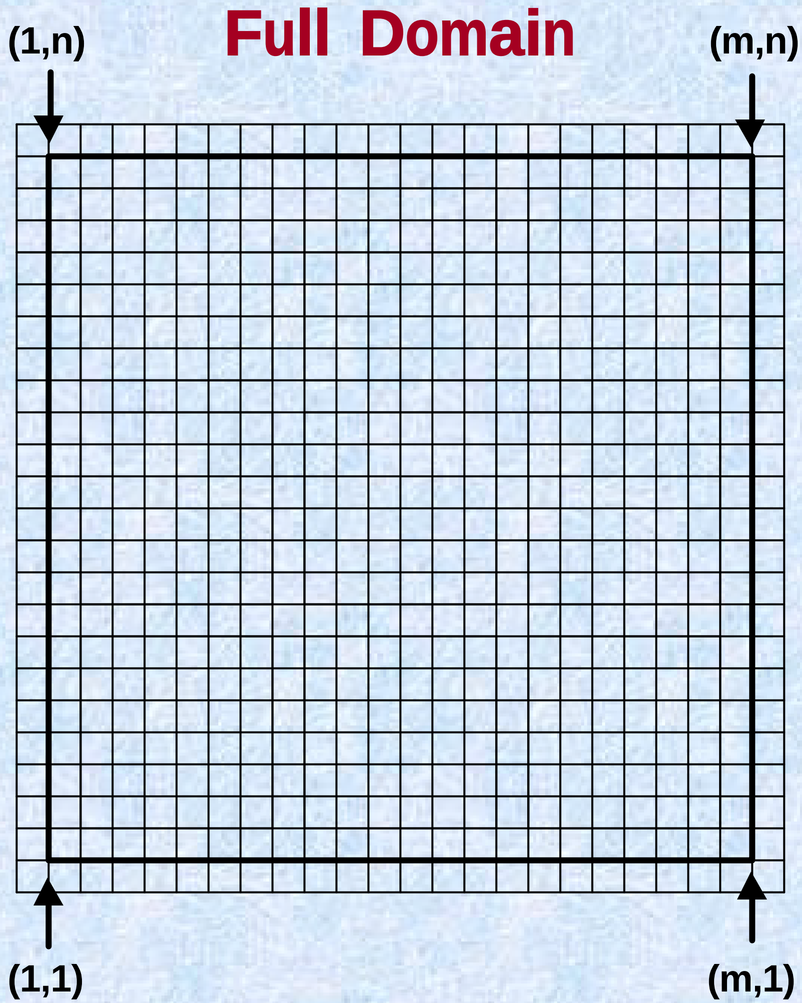
- **Scalar Code:**

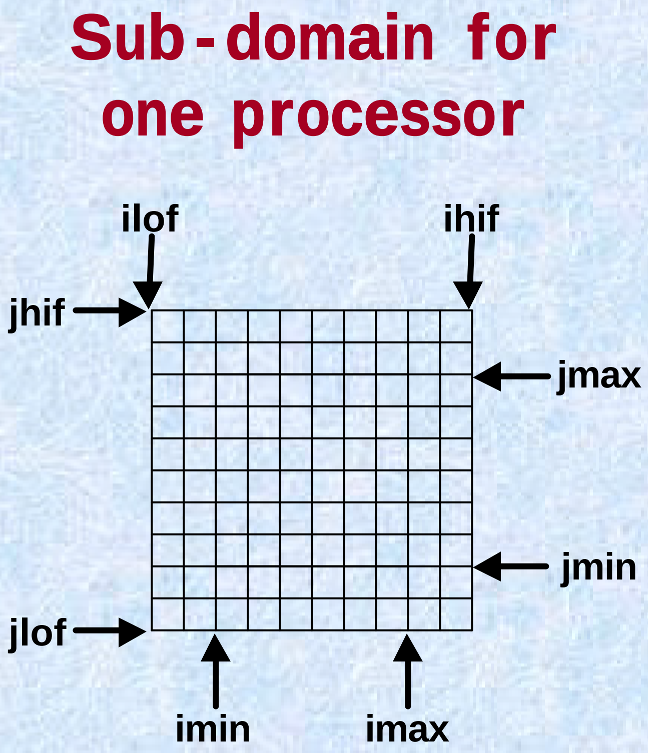
```
do k=1,kk
    do j=1,n
        do i=1,m
            a(i,j,k)=b(i,j,k)*2.0
        enddo
    enddo
enddo
```

- **Vector Code:**

```
do k=1,kk
    do i=1,m*n
        a(i,1,k)=b(i,1,k)*2.0
    enddo
enddo
```

COAMPS Domain Decomposition Using 2 Halo Rows





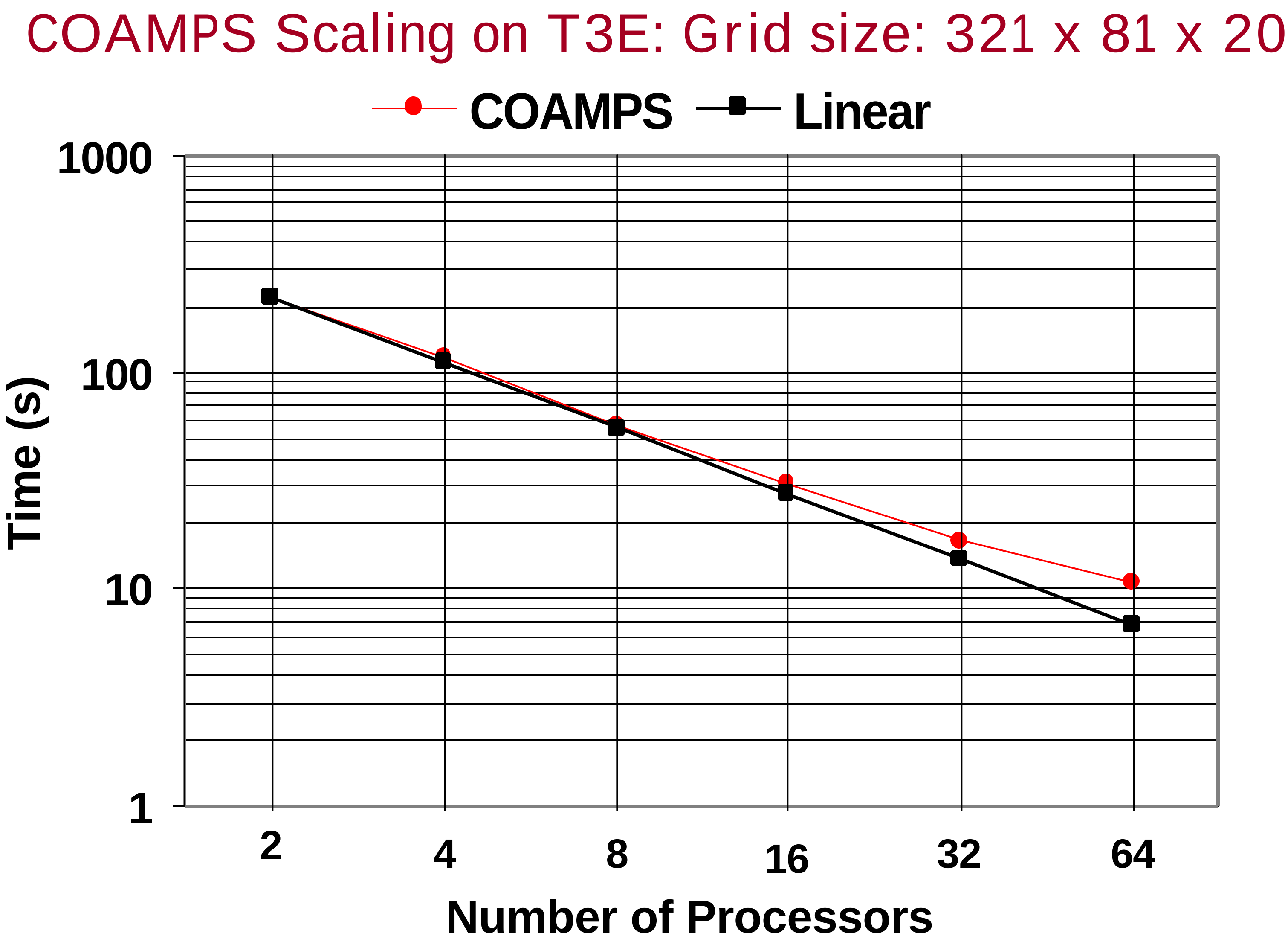
Scalable COAMPS

Initial Tests

- **Idealized Cases:**
  - Dry thermal bubble
  - Moist baroclinic wave development
  - 72 hour forecasts reproduces c90 results exactly
- **Real Data Cases:**
  - Forecast for individual cases reproduce c90 results exactly
  - Wall time using 40-processor SGI o2k is 57% of wall-time using 15-processor c90; o3k reduces running time an additional 33%
  - Data assimilation:
    - 2 week period
    - Minor differences due to:
      - Different interpolation methods
      - Different filters

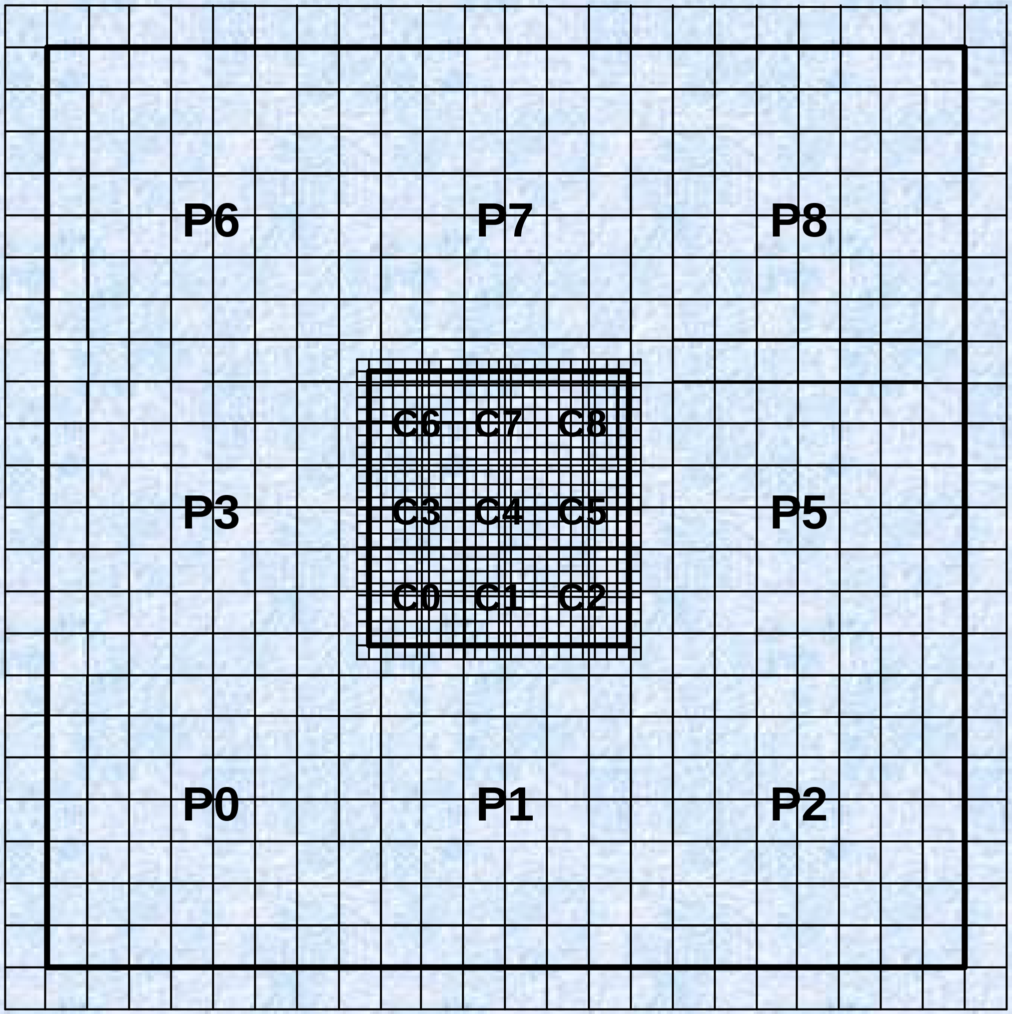
COAMPS

Scaling for idealized baroclinic wave development case



COAMPS Domain Decomposition for Two Nests

P: Parent (Coarse Mesh) Processors, **C:** Child (Fine Mesh) Processors



- Boundary conditions for C0 come from P0, P1, and P3
- Boundary conditions for C1 come from P1
- Boundary conditions for C2 come from P1, P2, and P5
- These communication rules become much more complicated when the child mesh is not so perfectly aligned with the parent mesh. In general, this is nearly always the case.

COAMPS MPI Moving Nest Software Development

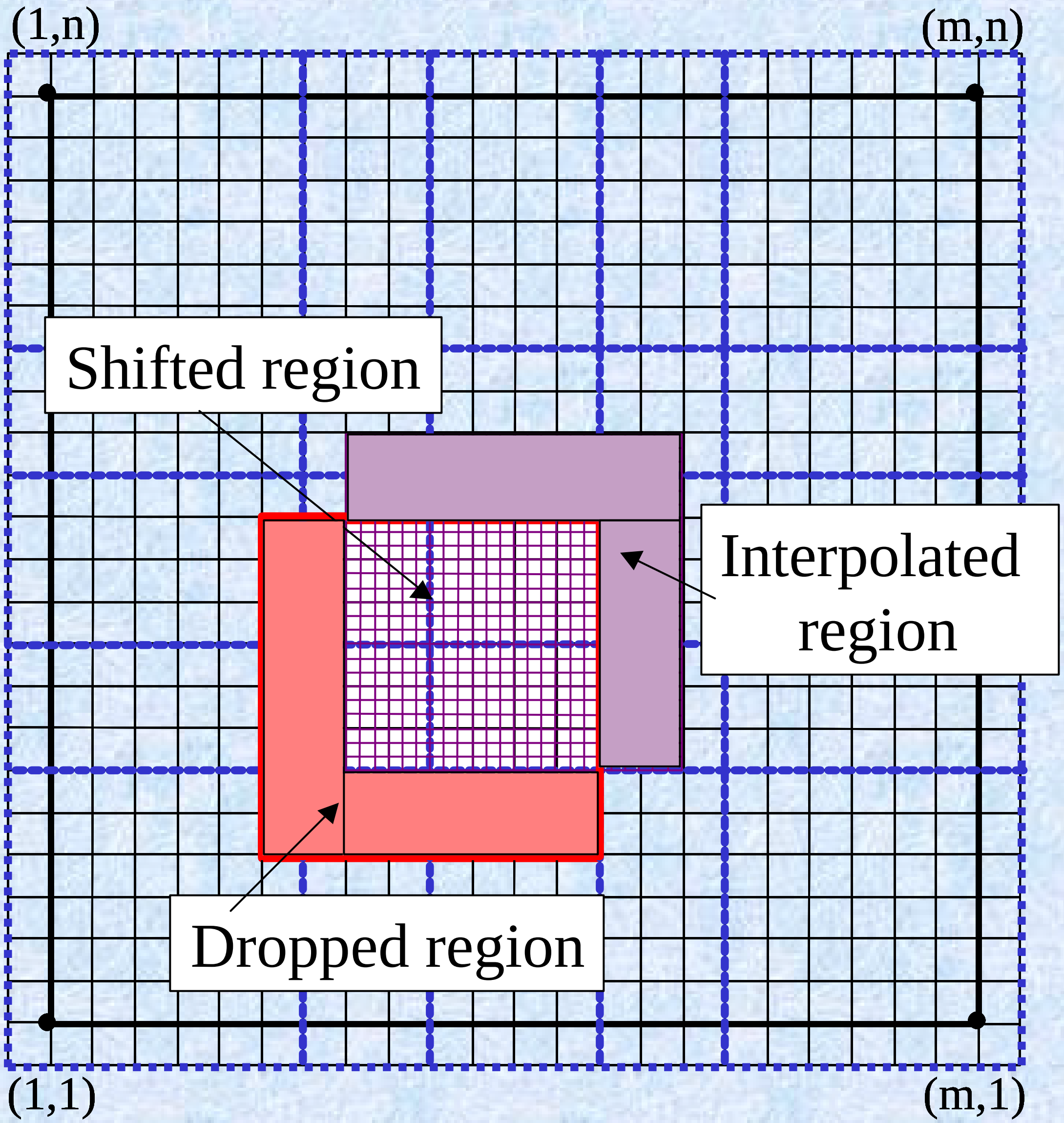
- **Software developed using MPI**
- **Makes use of existing COAMPS nesting software**
- **Advantages:**
  - **Allows for smaller nests (less resources required)**
  - **Flexibility in movement of nests:**
    - Namelist specified options:
      - Battle group option ("target" times/locations)
      - User specified grid point movement
    - Nests automatically move together
    - Automated tropical cyclone movement option (under development)

COAMPS MPI Moving Nest Software Development

Fixed Nest 1:
(m x n) points
3 x 3 domain decomposition
2 Halo Points

Moveable Nest 2:
Time = t0
Time = t1

MPI communications needed for shifted and interpolated areas

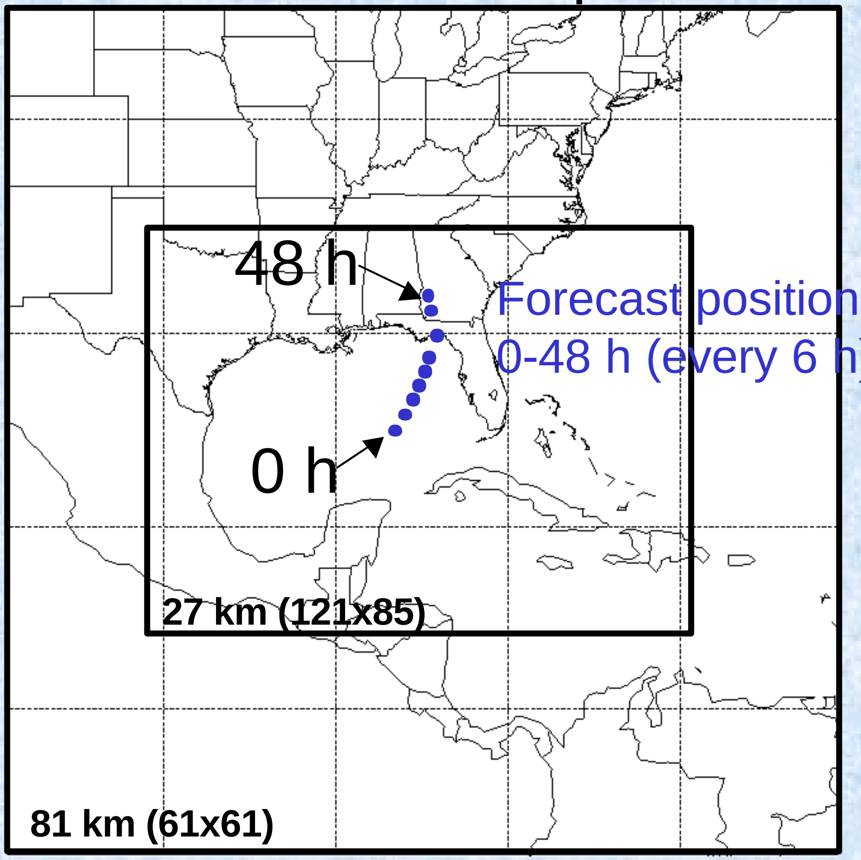


COAMPS MPI Moving Nests

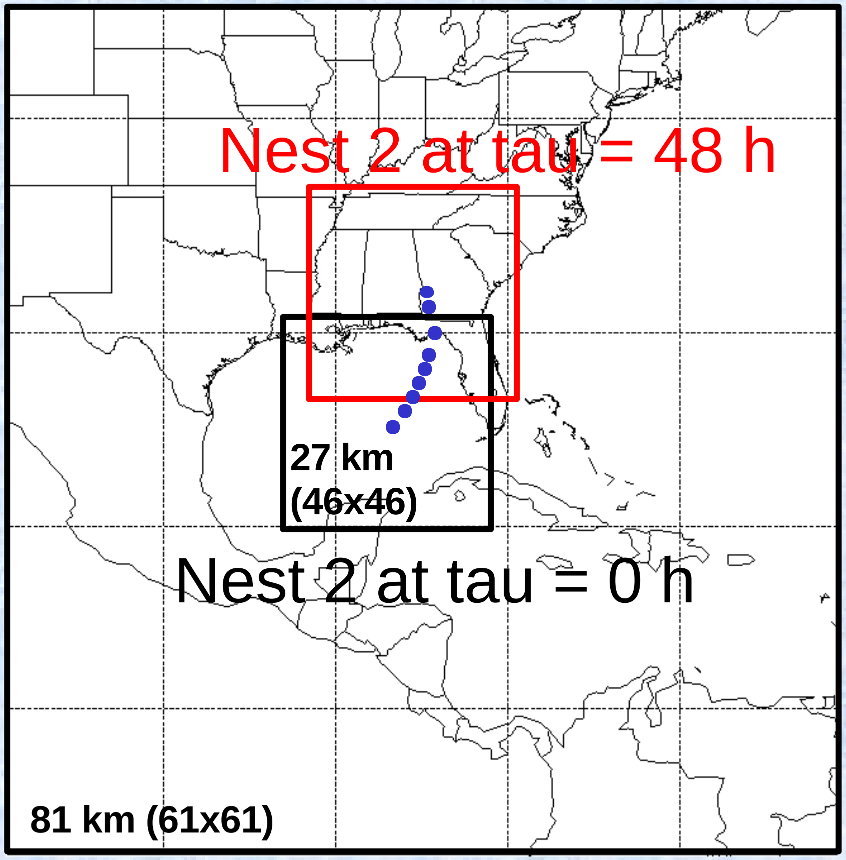
Hurricane Gordon 00Z September 17- 00Z September 19, 2000

Moving Nest Option is 2.7x Faster on O2K

Fixed Nest Option

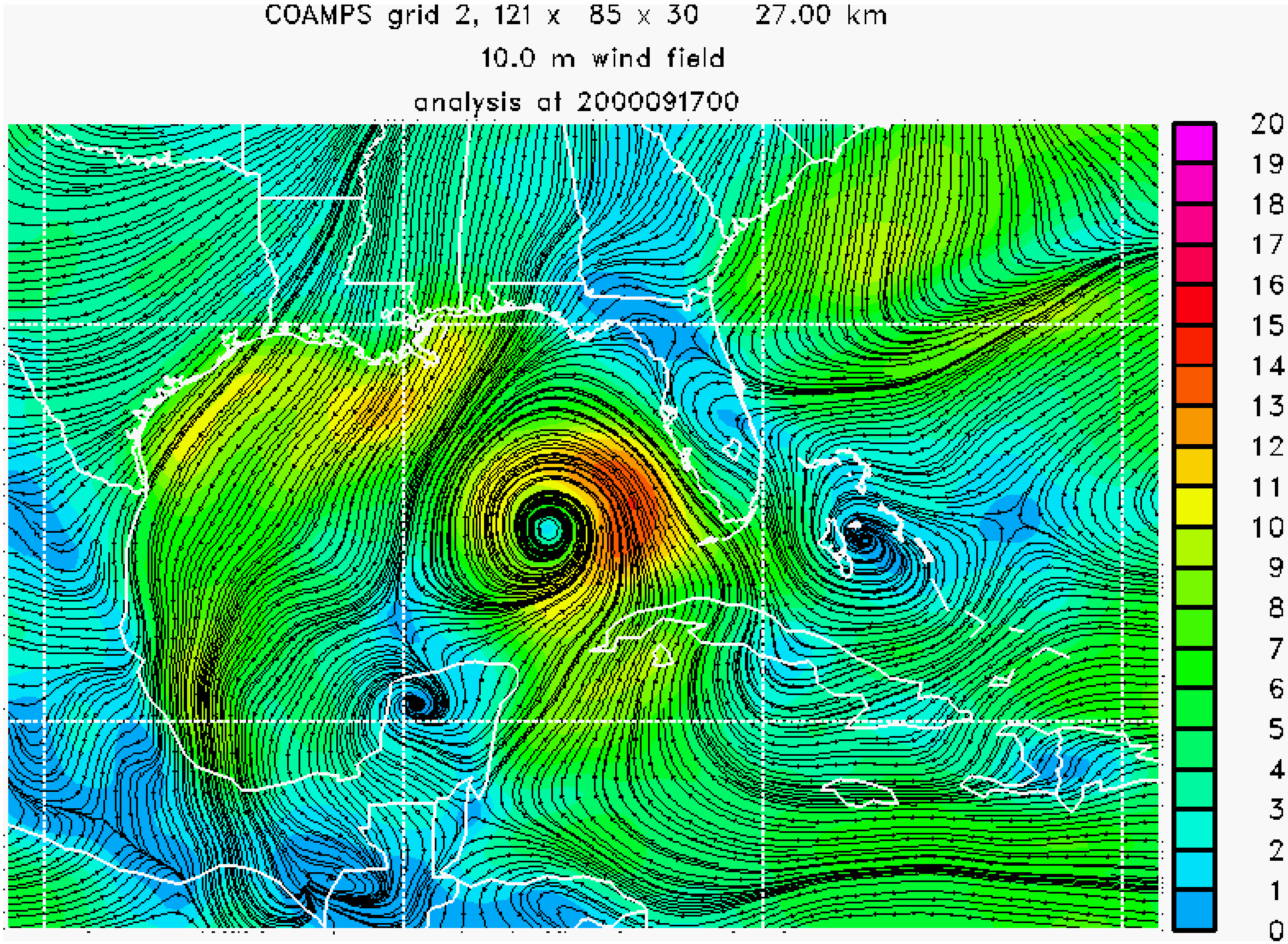


Moving Nest Option



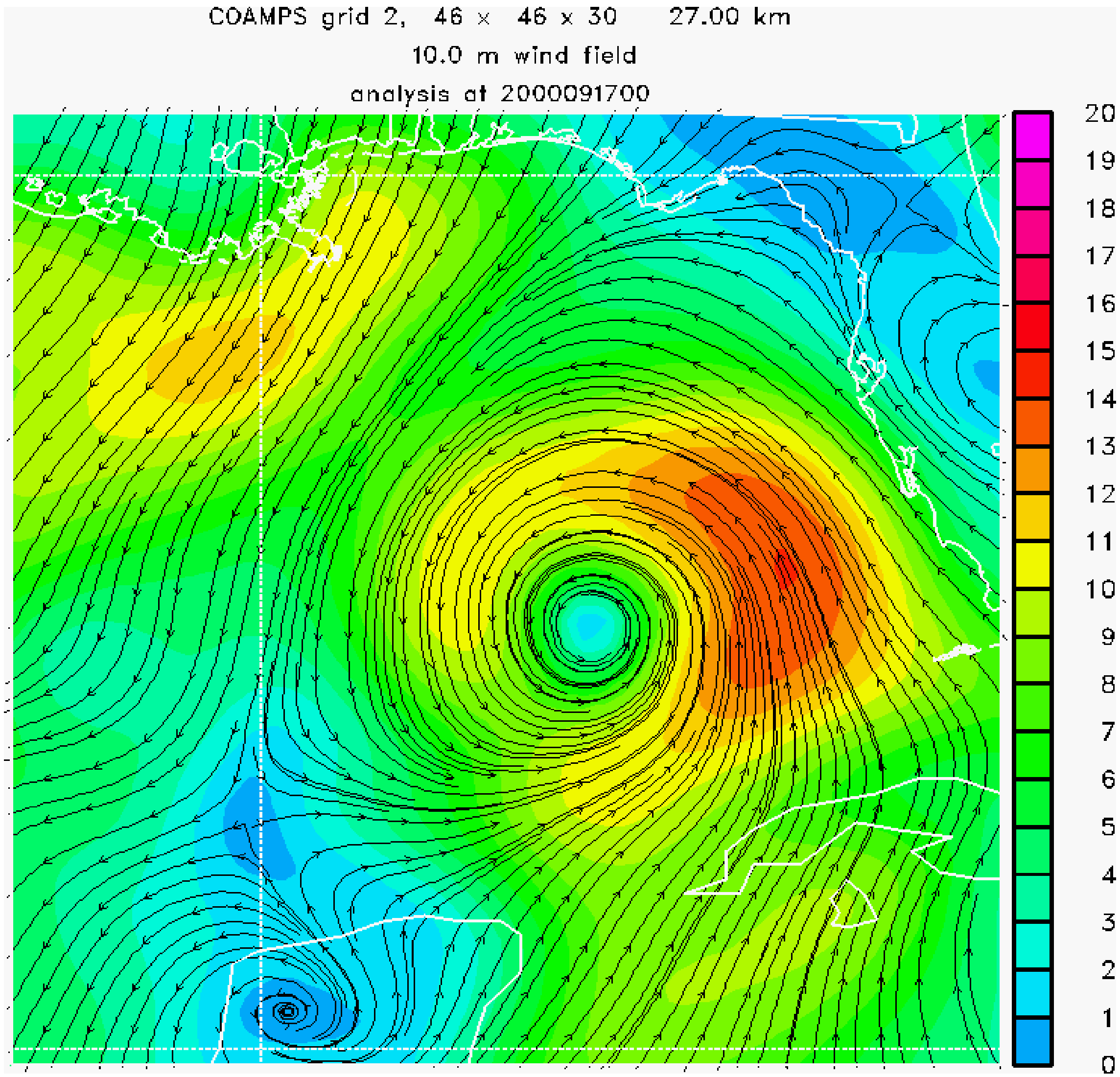
COAMPS Fixed Nest Animation

Hurricane Gordon



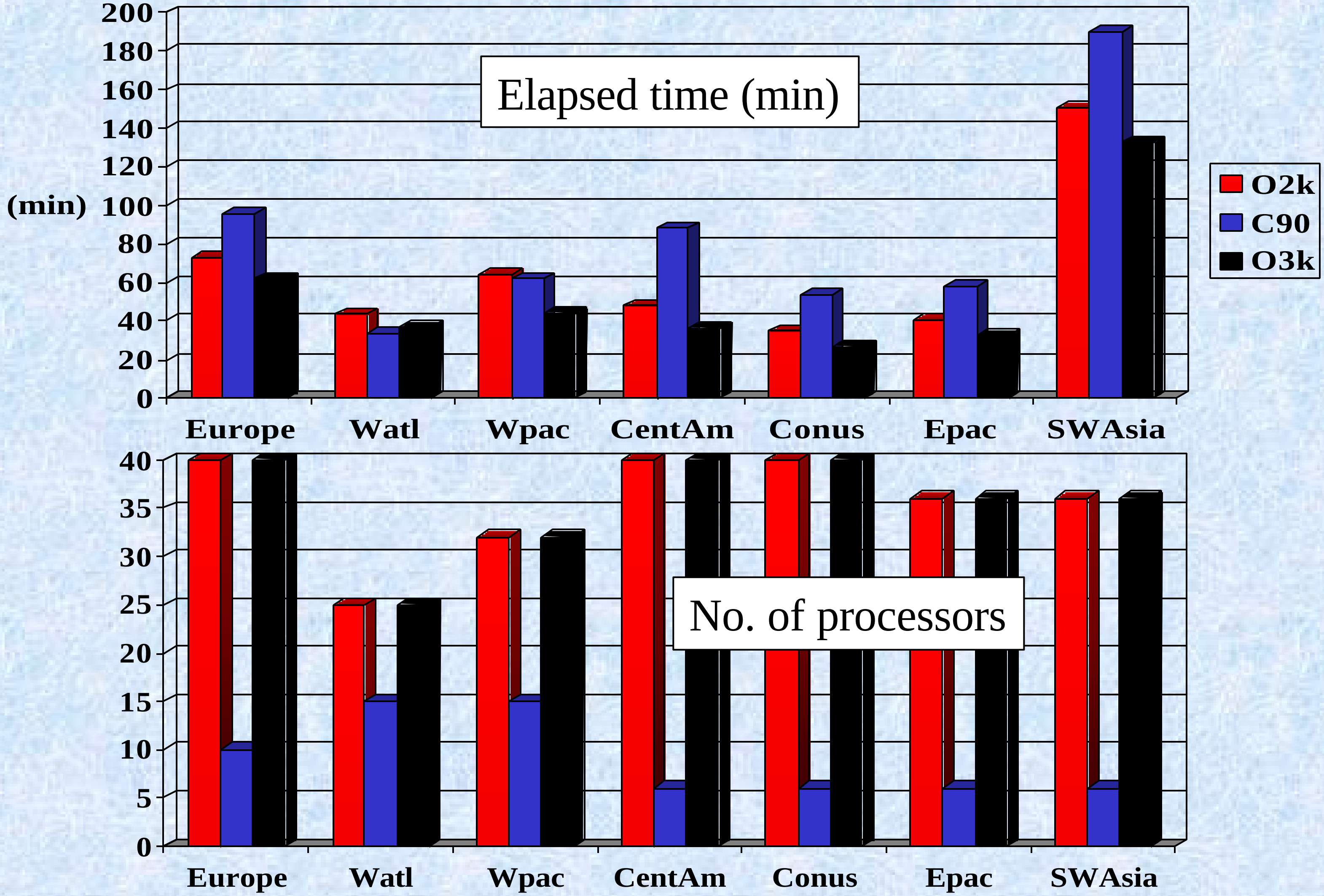
COAMPS Moving Nest Animation

Hurricane Gordon



Comparison of COAMPS on Cray C90 and SGI O2K

COAMPS 24 h forecasts in an operational/cpuset environment



- O2k elapsed times for 25-40 processors are comparable to C90 for 6-15 processors
- O3k reduces O2k elapsed times by 15-32%

Data courtesy of SGI Analyst Ken Taylor

Conclusion

Development and Performance of a Scalable Version of a Nonhydrostatic Model

- **FNMOC/HPC/LLNL moving to scalable architectures**
- **Developed scalable version of COAMPS:**
    - Successful NRL-LLNL collaboration
    - MPI and OpenMP use
    - x-, y- domain decomposition
    - Arbitrary number of halo points
    - Retains options of the shared memory version
    - Allows moving nested grids
- **Performance of scalable code:**
    - Demonstrated scaling to 60 processors, will test for > 60
    - Outperforms Cray c90/t90
    - Reproduces results of shared memory version
- **Shared memory version of COAMPS is frozen**
- **Scalable code being used for R&D and operations**

Future Plans

- **In Progress:**
  - Efficiency/Optimization:
    - Improved cache utilization
    - Examination of load imbalance
    - MPI-2 communications (one-way sends)
    - Test vectorization capabilities
  - Validation:
    - Different configurations
    - OpenMP/MPI comparisons across processors
  - Implementation of 3D variational analysis: NRL Atmospheric Variational Data Assimilation System (NAVDAS)
- **Scalable code in beta-ops at FNMOC**